2T

УДК 004.93+519.2

В.А. Таянов
Фізико-механічний інститут ім. Г.В. Карпенка НАН України, м. Львів
vtayanov@ipm.lviv.ua

Верхня та нижня оцінки помилок роботи узагальнених *k*NN класифікаторів

Пропонується підхід обчислення верхніх оцінок імовірності розпізнавання, що дає можливість використовувати їх для більш широкого класу моделей. Оцінки стосуються визначення стійкості покриття об'єктів класифікуючими алгоритмами на основі розподілу відстаней між об'єктами.

Вступ

На сьогоднішній день методи оцінки ймовірності правильного розпізнавання базуються на алгоритмах ковзаючого контролю (cross validation) [1]. Однак такі алгоритми ковзаючого контролю, як виключення по одному та ряд інших, вимагають великої кількості обчислень і комбінаторних перегрупувань вибірки. Тому потрібно розробити підходи щодо побудови верхніх оцінок для значно меншої кількості комбінаторних перегрупувань. Це можливо зробити тому, що навчальні дані практично завжди містять надлишок інформації, який проявляється в її частковому дублюванні. З точки зору перенавчання побудова верхніх оцінок означає, що перед алгоритмом класифікації поставлено найбільш складну задачу (подано найбільш складну вибірку), яка включає в себе довільні більш прості. Тобто верхні оцінки ймовірності правильного розпізнавання моделюють класифікацію найбільш незручних об'єктів навчальної вибірки. Якщо ймовірність того, що знайдуться більш несприятливі об'єкти на контрольній вибірці є малою, то доцільно говорити про те, що доповнення до цієї ймовірності визначатиме надійність оцінок. Поряд з тим, що будуються верхні оцінки для всієї вибірки або для сукупності підвибірок, також оцінюється зверху ймовірнісна стійкість покриття кожного об'єкта зокрема. Таким чином отримується більш точна та повна оцінка зверху для ймовірності правильного розпізнавання.

Важливі задачі теорії машинного навчання

В сучасній теорії машинного навчання існують дві серйозні проблеми: отримання точних верхніх оцінок імовірності такого негативного явища, як перенавчання, та способів боротьби з ним. На даний момент найбільш точні з відомих оцінок значно завищені. Експериментально вдалося встановити основні причини завищення оцінок. У порядку зменшення впливу вони є наступними:

– нехтування ефектом розшарування або локалізації сімейства алгоритмів. Дана проблема обумовлюється тим, що залежно від виду задачі використовується не все сімейство алгоритмів, а лише певна його частина. Коефіцієнт завищеності знаходиться в межах від декількох десятків до сотень тисяч;

– нехтування схожістю алгоритмів. Коефіцієнт завищеності становить для цього фактора від декількох сотень до десятків тисяч. Цей фактор завжди присутній і менш залежний від виду задачі, ніж перший;

– експоненційна апроксимація «хвоста» гіпергеометричного розподілу. В цьому випадку коефіцієнт завищеності може складати декілька десятків;
– верхня оцінка профілю різноманітності представляється одним скалярним коефіцієнтом різноманітності. Коефіцієнт завищеності часто близько одиниці, однак у деяких випадках може досягати декількох десятків.

Причина ефекту перенавчання обумовлюється тим, що використовуються алгоритми з мінімальним числом помилок на навчальній вибірці, тобто відбувається однобічне налаштування цих алгоритмів. Перенавчання буде тим більшим, чим більша композиція алгоритмів використовується для класифікації, якщо ці алгоритми беруться з розподілу випадково і незалежно. У випадку залежності алгоритмів (в реальній ситуації вони, як правило, такими і є) перенавчання зменшиться. Отже, при виборі навіть одного з двох алгоритмів може виникнути перенавчання. Розшарування алгоритмів за числом помилок та збільшення їхньої подібності зменшують імовірність перенавчання. Розглянемо для прикладу дуплет «вибірка-алгоритм». Кожний алгоритм покриває певну частину об’єктів навчальної вибірки. Якщо використовувати внутрішні критерії [2] (наприклад, у випадку метричних класифікаторів), то можна оцінити стійкість цього покриття і звузити число покритих об’єктів згідно із заданим рівнем стійкості. Таким чином, для того щоб покрити більшу кількість об’єктів, потрібно застосувати більшу кількість алгоритмів. Ці алгоритми мають бути схожими і мати різний рівень помилок. Однак, якщо використовуються тестові дані, до яких композиція алгоритмів неадаптована, то помилка класифікації може досить помітно відрізнятися від мінімальної, отриманої на навчальних даних.

Побудова оцінок імовірнісної стійкості покриття об’єктів алгоритмами типу *k*NN для одиночних випробувань

Якість роботи класифікаторів, що будуються на основі рангового голосування та з використанням розділювальних гіперплощин (R моделей [3]) прийнято характеризувати через поняття відступу (margin), що представляє відстань об’єкта від розділювальної гіперплощини. Чим більший відступ, тим кращим вважається класифікатор. Однак якщо всі об’єкти або переважна їх більшість мають приблизно однаковий відступ і групуються один біля одного, то в цьому випадку різко падає їх інформативність. Це означає, що замість всіх об’єктів можна залишити один або декілька, що використовуються для навчання. Такий підхід породжує одну з головних причин, що обумовлюють ефект перенавчання. Однобічне налаштування алгоритму на основі близької за суттю навчальної інформації призводить до того, що на контрольній вибірці він може часто помилятись, навіть якщо не помилявся на навчальній вибірці. Дійсно, ймовірність того, що в умовах навчальної вибірки зустрінеться така ж ситуація, є близькою до нуля.

Тому для навчання прийнято використовувати несхожі і «важкі» для алгоритму об’єкти з малими значеннями відступу. Ця ідея використана, зокрема, у методі опорних векторів (Support Vector Machine) або методі зваженого голосування. Застосуємо узагальнений підхід для характеристики класифікаторів на основі поняття відступу. Результатом роботи метричних класифікаторів є ранжовані дані (посортовані за ступенем подібності до тестового об’єкти бази даних). Для таких класифікаторів поняття відступу представляється наступним чином. Вводиться еквівалентна до класичного

відступу характеристика, яка для даного об'єкта може бути представлена як відносна відстань між його відстанями від тестового об'єкта та від усередненого об'єкта бази даних або останнього об'єкта з однорідної (стратегічної) [4] послідовності «своїх» об'єктів. Передбачається, що хоча б частина «своїх» об'єктів розташовується на початку списку можливих претендентів. Таким чином, гарантується коректність даного означення.

Для більш точного означення даної характеристики потрібно ввести поняття розподілу відстаней між об'єктами. Оскільки значення відстаней може бути довільним, то процедура непараметричного оцінювання розподілу неусіченими ядерними функціями буде коректною.

Нехай непараметрично оцінена густина розподілу відстаней між об'єктами, заданих векторами $\mathbf{x}$ та $\mathbf{y}$: $p(x)$, $x \to d(\mathbf{x},\mathbf{y})$. Згідно з нерівністю Чебишева [5], ймовірність того, що знайдеться відстань, яка перевищить деяке порогове значення відстаней θ, дорівнює $\int\limits_{|x|\geq\theta} p(x)dx \leq \frac{\sigma^2}{\theta^2}$.

Розглянемо випадок рівності математичного сподівання та моди розподілу $p(x)$. Верхня межа одномодального розподілу з модою $\mu = 0$ за допомогою нерівності Гауса [6] представляється у вигляді:

$$P(|x-\mu| \geq \lambda\tau) \leq \frac{4}{9\lambda^2}, \tag{1}$$

де $\tau^2 \equiv \sigma^2 + (\mu - \mu_0)^2$.

Нехай $\mu = \mu_0 = 0$ і $\tau \equiv \sigma$. Тоді поріг $\theta = \lambda\tau = \lambda\sigma$, а $\lambda = \frac{\theta}{\sigma}$. Отже, нерівність Гауса для порогу θ може бути представлена у вигляді:

$$\int\limits_{|x|\geq\theta} p(x)dx \leq \frac{4\sigma^2}{9\theta^2}. \tag{2}$$

Таким чином, згідно з нерівністю Гауса для одномодальних розподілів з модою, що дорівнює математичному сподіванню, оцінка є в 2,25 разів кращою за ту, яка отримується за допомогою нерівності Чебишева. Це максимально хороша оцінка за умови, що невідомий конкретний вид розподілу, а відомі лише певні його властивості. Необхідною і достатньою умовою рівності моди математичному сподіванню є симетричність одномодального розподілу. Однак у загальному випадку реальний закон розподілу не є симетричним. При цьому можливі ліва або права асиметрії функції густини розподілу ймовірностей (ФГРЙ).

Побудова оцінок імовірнісної стійкості покриття об'єктів алгоритмами типу *k*NN для певних класів розподілів відстаней між об'єктами

Розділимо ФГРЙ на дві частини, а саме: частину, що знаходиться справа та зліва від максимуму. Якщо права частина ФГРЙ більша за ліву, то вважається, що це права асиметрія, а якщо навпаки – то ліва (рис. 1, 2). Розглянемо оцінки за допомогою нерівності Гауса для обох випадків. У випадку правої асиметрії застосуємо наступний прийом. Зробимо розподіл симетричним за лівою частиною, тобто ліву частину залишаємо незмінною та відображаємо її симетрично відносно максимуму замість вихідної

правої частини. Нехай деяка точка x_0 належить лівій частині розподілу. Тоді функція розподілу ймовірностей (ФРЙ) $P(X < x_0)$ для симетричного випадку буде завжди більшою від вихідного значення у кожній точці лівої частини розподілу. Відзначимо, що нас цікавлять перші об'єкти у списку можливих претендентів, що відповідають лівій частині розподілу. Тоді ФРЙ буде верхньою оцінкою для помилки розпізнавання.

Проаналізуємо отриманий результат. Попередньо відзначимо, що для кращого розуміння прийому, а також спрощеної інтерпретації результатів немає необхідності у нормуванні ФГРЙ до одиничної площі. Фактичне зменшення площі під кривою ФГРЙ означає, що результати розпізнавання і прийняття рішення впливають не на всі об'єкти, а лише на їх частину, що відповідає реальній ситуації. До того ж інтерес представляють відхилення відстаней вліво від математичного сподівання при використанні *k*NN класифікаторів з невеликими значеннями *k*. Оскільки оцінка дисперсії ФГРЙ для побудови оцінки Гауса проводилась за лівою частиною, то очевидно, що ця оцінка у випадку симетричної ФГРЙ буде меншою за вихідну, що робить її більш точною. До того ж симетрія дозволяє зробити оцінку Гауса максимально точною згідно з нерівністю (2), а все разом дозволяє суттєво покращити загальну верхню оцінку.

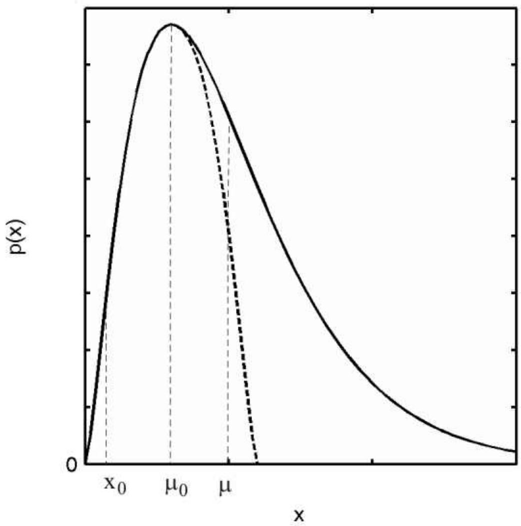


Рисунок 1 – Права асиметрія ФГРЙ

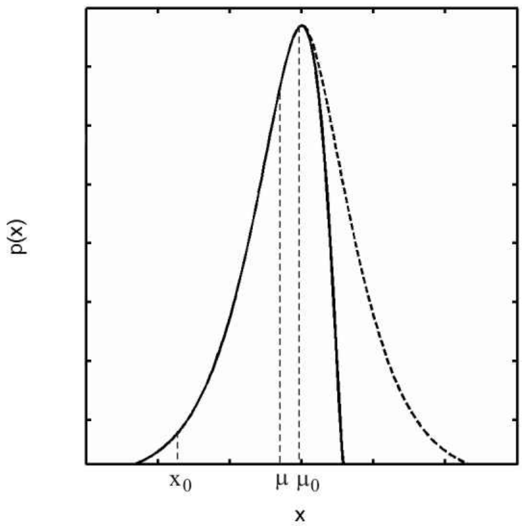


Рисунок 2 – Ліва асиметрія ФГРЙ

Розглянемо ФГРЙ у випадку лівої асиметрії. При цьому єдино можливим є лише симетричне відображення лівої частини у праву, оскільки лише тоді можлива коректна верхня оцінка. Тепер дисперсія симетричної ФГРЙ буде більшою, ніж вихідної, а єдиною перевагою такого перетворення буде симетричність щойно отриманого закону розподілу.

У даному випадку також немає необхідності у нормуванні ФГРЙ. Збільшення площі під кривою означає, що включені додаткові об'єкти, які не приймають участі у розпізнаванні. Це погіршує оцінку Гауса, оскільки збільшується значення оціненої дисперсії. Рішення про те, яку оцінку використовувати – з перетворенням симетрії або по вихідному розподілу – необхідно приймати, маючи значення математичного сподівання, моди та дисперсії обох розподілів.

Проаналізуємо зв'язок оцінки Гауса зі значенням ФРЙ $P(X < x_0)$. Права частина ФГРЙ не представляє інтересу, тому якщо замість оцінки Гауса взяти ФРЙ, то це буде оцінкою зверху стосовно самої оцінки. При цьому не мають значення ані вид асиметрії, ані сама асиметрія в законі розподілу взагалі. Тобто верхня оцінка значеннями

ФРЙ щодо до оцінки Гауса стосується як симетричних, так і асиметричних законів розподілу. Таким чином, завищеність оцінки Гауса щодо значень ФРЙ компенсується лише у випадку правої асиметрії. У випадку лівої асиметрії ступінь компенсації залежить від співвідношення між значенням дисперсії та різниці $|\mu - \mu_0|$.

Якщо ФГРЙ не має чітко вираженої структури (існування максимуму, симетрія), тоді можна скористатися непараметричним оцінюванням, в результаті якого отримується неперервна ФГРЙ. Цю функцію можна інтегрувати та диференціювати. З іншого боку, як показано в [7], при збільшенні розміру чужого класу та незмінному розмірі свого зменшується ймовірність правильного розпізнавання. Тому потрібно по можливості забезпечувати рівність розмірів класів, щоб не було перекосу в пріоритетах. Якщо ж цього не вдасться зробити, то при обчисленні слід враховувати ймовірність появи об'єктів того чи іншого класу.

Оскільки нормальна ФГРЙ характеризується мінімальною помилкою класифікації для даного порогу θ [8] і не перевищує $\frac{4\sigma^2}{9\theta^2}$ у випадку одномодальної симетричної ФГРЙ або ФГРЙ з правою асиметрією, то двостороння нерівність для даної помилки розпізнавання ε може бути записана у вигляді:

$$0{,}5(1 - erf(\frac{\theta}{\sigma})) \le \varepsilon \le \frac{4\sigma^2}{9\theta^2}, \qquad (3)$$

де $\mu = 0$.

Проаналізуємо можливу загальну форму потенційно отримуваних ФГРЙ відстаней між об'єктами. Всі розподіли матимуть максимуми, оскільки функція ФГРЙ існує на інтервалі $[0,\infty)$, а густина в околі 0 та для великих відстаней не може бути значною, тому що такі події малоймовірні. Права асиметрія є набагато більш імовірною, оскільки розподіл відстаней обмежений з лівого боку нулем, а з правого боку він не має строгих обмежень.

Оцінки ймовірнісної стійкості покриття об'єктів алгоритмами типу *k*NN в умовах двох класів, що мають задані розміри

Розглянемо поширену задачу класифікації в умовах двох класів. Позначимо розміри класів як s_1 та s_2. Тоді, якщо ймовірність заміщення об'єкта з класу розміром s_1 у межах довірчого інтервалу дорівнює ε_1, то ймовірність незаміщення об'єктів із цього класу об'єктами з класу розміром s_2 дорівнює $(1-\varepsilon_1)^{s_2}$ за умови незалежності об'єктів [7]. Для іншого класу при відповідних змінах у позначеннях ця ймовірність дорівнюватиме $(1-\varepsilon_2)^{s_1}$. Якщо тепер ввести деякий віртуальний клас і вважати, що заміщення якогось об'єкта цього класу об'єктами із згаданих двох класів є вірогідною подією, то можна записати наступне рівняння:

$$\gamma((1-\varepsilon_1)^{s_2} + (1-\varepsilon_2)^{s_1}) = 1, \qquad (4)$$

звідки множник пропорційності γ знаходиться тривіально.

Часом зустрічаються ситуації, коли відстані між об'єктами дорівнюють 0. При цьому непараметричний розподіл одного з класів може мати максимум у точці, що відповідає нульовій відстані. Нехай густини розподілів у нульовій точці дорівнюють $p_1(0)$. та $p_2(0)$. Оцінка співвідношення між імовірностями може бути задана у вигляді

$\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$ або $\ln\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$. При цьому потрібно зробити граничний перехід від ФРЙ до ФГРЙ, оскільки вони пов'язані між собою операцією диференціювання. Співвідношення $\ln\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$ ($\ln\frac{p_2(0)^{s_1}}{p_1(0)^{s_2}}$) або в загальному $\ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}}$ ($\ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}}$) можна використати для побудови класифікатора виду

$$\begin{matrix} \ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} > \gamma_1; & & \ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} > \gamma_2; \\ & \text{чи} & \\ \ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} < \gamma_1, & & \ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} < \gamma_2, \end{matrix} \qquad (5)$$

де значення $\ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} = 0$ або $\ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} = 0$ не впливають на результати класифікації і рішення може бути прийняте на користь довільного класу. У випадку непараметричного оцінювання ймовірність такого значення практично дорівнює 0.

Висновки

В роботі побудовані та досліджені оцінки ймовірності правильної класифікації для класифікаторів, що використовують як міру подібності функцію відстані. Результати оцінювання проводились на основі функції розподілу відстаней між об'єктами. При цьому розглянуті різні часткові випадку функції розподілу за формою. Побудовані двосторонні верхні оцінки у випадку одиночних випробувань та для заданих розмірів двох класів. Запропонований метод класифікації на основі співвідношення густин розподілу ймовірностей у нульовій та довільній точках.

Література

1. Vorontsov K.V. Combinatorial probability and the tightness of generalization bounds / K.V. Vorontsov // Pattern Recognition and Image Analysis. – 2008. – Vol. 18. – № 2. – P. 243-259.
2. Kapustii B.E. Classifier optimization in small sample size condition / B.E. Kapustii, B.P. Rusyn, and V.A. Tayanov // Automatic Control and Computer Sciences. – 2006, 40 (5). – P. 17-22.
3. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации / Ю.И. Журавлев // Проблемы кибернетики. – 1978. – Т. 33. – С. 5-68.
4. Капустій Б.О. Комбінаторна оцінка впливу зменшення інформаційного покриття класів на узагальнюючу властивість 1NN алгоритмів класифікації / Б.О. Капустій, Б.П. Русин, В.А. Таянов // Искусственный интеллект. – 2008. – № 1. – С. 49-54.
5. Weisstein E.W. Chebyshev Inequality, From MathWorld [Електронний ресурс] / E.W. Weisstein // A Wolfram Web Resource. – Режим доступу : http://mathworld.wolfram.com / ChebyshevInequality.html, 10.12.2008.
6. Weisstein E.W. Gauss's Inequality, From MathWorld [Електронний ресурс] / E.W. Weisstein // A Wolfram Web Resource. – Режим доступу : http://mathworld.wolfram.com/GausssInequality.html, 10.12.2008.
7. Капустій Б.О. Новый подход к определению вероятности правильного распознавания объектов множеств / Б.О. Капустій, Б.П. Русин, В.А. Таянов // УСиМ. – 2005. – № 2. – С. 8-13.
8. Ту Дж. Принципы распознавания образов / Дж. Ту, Р. Гонсалес– М. : Мир, 1978. – 416 с.

В.А. Таянов

Верхняя и нижняя оценки ошибок работы обобщённых *k*NN классификаторов

Предлагается подход вычисления верхних оценок вероятности правильного распознавания, что дает возможность использовать их для более широкого класса моделей. Оценки касаются определения устойчивости покрытия объектов классифицирующими алгоритмами на основании распределения расстояний между объектами.

Стаття надійшла до редакції 02.04.2009.